

Guaranteed Model-Based Fault Detection in Cyber-Physical Systems: A Model Invalidation Approach

Farshad Harirchi and Necmiye Ozay, ^{*}

September 21, 2016

Abstract

This paper presents a sound and complete fault detection approach for cyber-physical systems represented by hidden-mode switched affine models with time varying parametric uncertainty. The fault detection approach builds upon techniques from model invalidation. In particular, a set-membership approach is taken where the noisy input-output data is compared to the set of behaviors of a nominal model. As we show, this set-membership check can be reduced to the feasibility of a mixed-integer linear programming (MILP) problem, which can be solved efficiently by leveraging the state-of-the-art MILP solvers. In the second part of the paper, given a system model and a fault model, the concept of T -detectability is introduced. If a pair of system and fault models satisfies T -detectability property for a finite T , this allows the model invalidation algorithm to be implemented in a receding horizon manner, without compromising detection guarantees. In addition, the concept of weak-detectability is introduced which extends the proposed approach to a more expressive class of fault models that capture language constraints on the mode sequences. Finally, the efficiency of the approach is illustrated with numerical examples motivated by smart building radiant systems.

1 Introduction

Cyber-physical systems are combinations of physical processes and embedded computers. The embedded computers collect data from the process

^{*}This work is supported in part by DARPA grant N66001-14-1-4045.

[†]The authors are with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109. {harirchi,necmiye}@umich.edu

through sensors and control it in a closed-loop manner. With the increase in data acquisition and storage capacity and the decrease in sensor costs, it is possible to collect large amounts of data during the operation of complex cyber-physical systems. For instance, “a four-engine jumbo jet can create 640 terabytes of data in just one crossing of the Atlantic Ocean” [1]. As discussed in [2], this exponential growth in the data collection capabilities is a major challenge for systems and control community. Sensor/information-rich networked cyber-physical systems, from air traffic or energy networks to smart buildings, are getting tightly integrated into our daily lives. As such, their safety-criticality increases. For such systems, it is crucial to detect faults or anomalies in real-time to support the decision-making process and to prevent potential large-scale failures.

1.1 Contributions

This paper presents a fault detection scheme to enhance the reliability of a class of cyber-physical systems that are represented by hidden-mode switched affine models with time-varying parametric uncertainty. This modeling framework is quite expressive and can be used to describe a wide range of cyber-physical systems such as heated ventilation and air conditioning (HVAC) systems in smart buildings [3], wind turbines [4], power systems and power electronics, automotive systems, aircrafts, air traffic, and network and congestion [5]. We model faults also in this framework allowing us to capture many scenarios including cascaded faults or various types of cyber or physical attacks. Note that linear time invariant systems with or without noise or affine parametric uncertainty are special cases of the modeling framework.

The proposed fault detection scheme builds on set-membership model invalidation approaches [6–8]. Unlike many set-membership methods that compute explicit set representations and propagate them via set-valued observers, the proposed method uses an online optimization formulation for model invalidation that keeps implicit constraints to represent sets. In particular, we show that model invalidation problem for this class of systems can be reduced to the feasibility of a MILP problem, which can be checked efficiently using state-of-the-art solvers [9]. Additionally, the concept of T -detectability is introduced, which enables us to apply the model invalidation approach for fault detection in a receding horizon manner (with a horizon size T) without losing detection guarantees. Even though there are some practical systems with fault models that are T -detectable, a limited class of system and fault models satisfy this property. We further discuss weak de-

tectability that incorporates language constraints on the switching sequences of the faults to enable detection of a broader class of faults. Algorithms that can be used to find the minimum T (if it exists) are presented.

A preliminary version of this paper is published in [10]. The current paper significantly extends the model class by allowing uncertainty in variables and considering language constraints in the mode sequence. Moreover, more efficient MILP-based necessary and sufficient conditions for verifying T -detectability is given and some connections to mode observability of switched systems are pointed out.

1.2 Literature Review on Fault Detection

Model-based fault detection has a long history starting with early work on failure detection filters [11, 12]. A vast majority of fault detection methods are based on residual generation, where the residual is evaluated by simple thresholding methods or more complicated classifiers to decide between faulty and normal behaviors [13–16]. The residual generation methods are classified into three main categories [13]: parameter estimation-based [17], observer-based [18–20] and parity equation-based [21] techniques. All these residual generation techniques can be implemented in real-time, but even when a specific fault model exists, their behavior is usually analyzed only asymptotically and they fail to provide any finite-time detection guarantees.

As an alternative to residual generation, set-membership fault detection methods have been proposed both for passive [6, 22] and active [23–25] fault detection. Some of these methods proceed by computing convex-hulls of potentially non-convex reachable sets of the system and comparing the actual output to this set [6]. Since, reachable sets are over-approximated, this leads to only sufficient conditions, that is, they guarantee there are no false alarms. Scott *et al.* [25] pose a mixed-integer quadratic programming problem to find an optimal separating input to detect faults. They use zonotopes to represent and propagate state constraints. The order of the zonotopes is used to trade-off between the complexity and the conservativeness of the approach. Despite the fact that all the above mentioned set-membership methods are proposed for linear systems, their scalability is somewhat limited to be applied in real-time [25].

Non-linear and hybrid systems have also attracted notable attention from fault detection community. Most of the research is concentrated around residual generation type methods [26–28]. De Persis and Isidori [29] develop analytical necessary and sufficient conditions under which the problem of fault detection and isolation becomes solvable for non-linear systems.

However, they do not particularly address the computational aspects of the problem. Observer-based methods are employed for fault diagnosis in hybrid systems both when the discrete mode is observed [30] or hidden [31], using variants of Kalman filters. In recent work [32], Deng *et al.* investigate fault diagnosis problem in hybrid systems, by constructing a finite abstraction for the hybrid automaton and analyzing the diagnosability of the abstract system using tools from discrete-event systems, providing some detection guarantees. However, the mode signal is assumed to be observed in [32].

The notion of T -detectability is closely related to distinguishability of dynamical systems [33] and observability in switched systems [34, 35]. The distinguishability checks if there exists a non-zero input which makes the trajectory of two linear time-invariant systems distinguishable or not. Lou *et. al* [36] introduced the concept of input-distinguishability, which restricts distinguishability to all non-zero inputs contained in a convex set. Rosa and Silvestre [37] consider the input-distinguishability for discrete time linear time-invariant systems with bounded disturbance and noise. They provide a necessary and sufficient rank condition to check for a given size of time horizon (T) if two systems are input distinguishable or not. This condition is seldom satisfied in practice, hence they added extra constraint which enforces the persistence of excitation for disturbance [37]. The T -detectability concept introduced here is closely related to absolute input-distinguishability [37], but provides necessary and sufficient conditions for the input-distinguishability of two hidden-mode switched affine models subject to process and measurement noise. The time horizon T in detectability is an upper bound on the detection delays (time from the occurrence of fault to detection alarm) [38–40].

Model invalidation was originally proposed as a way to build trust in models obtained through a system identification step or discard/improve them before using these models in robust control design [41]. In model invalidation problem, one starts with a family of models (i.e., a priori or admissible model set) and experimental input-output data collected from a system (i.e., a finite execution trace) and tries to determine whether the experimental data can be generated by one of the models in the initial model family. Its relation to fault and anomaly detection has been pointed out in [6–8, 42] for linear time-varying systems and hybrid systems in autoregressive form. A fault detection scheme based on model-invalidation for polynomial state space models subject to noise and uncertainty in parameters is recently proposed in [43]. The convex relaxations for model invalidation problem mostly provide sufficient conditions that can be efficiently checked to detect faults but only necessary for large relaxation orders.

Notation: Let $\mathbf{x} \in \mathbb{R}^n$ denote a vector and \mathbf{x}^i indicate its i^{th} element. Also, let $\mathbf{M} \in \mathbb{R}^{n \times m}$ represent a matrix and $\mathbf{M}^{i,j}$ indicate the element on the i^{th} row and j^{th} column of the matrix \mathbf{M} . The infinity norm of a vector \mathbf{x} is denoted by $\|\mathbf{x}\| \doteq \max_i \mathbf{x}^i$. The set of positive and non-negative integers up to n are denoted by \mathbb{Z}_n^+ and \mathbb{Z}_n^0 , respectively. The Hadamard product of two matrices \mathbf{M}_1 and \mathbf{M}_2 of the same dimensions is indicated by $\mathbf{M} = \mathbf{M}_1 \odot \mathbf{M}_2$, and defined as $\mathbf{M}^{i,j} \doteq \mathbf{M}_1^{i,j} \mathbf{M}_2^{i,j}$ for all i, j .

2 Modeling Framework

In this section, we describe the class of systems and the modeling framework considered in this paper. In particular, we consider discrete-time hidden-mode switched affine models with time-varying parametric uncertainty that are subject to noise. We refer to these models as SWA models. SWA models consist of a collection of affine models, which we define first.

Definition 1 (*Affine Model*) *An affine model G^Δ with time-varying parametric uncertainty has the following form:*

$$\begin{aligned} \mathbf{x}_{k+1} &= (\mathbf{A} + \mathbf{A}_k^\Delta) \mathbf{x}_k + (\mathbf{B} + \mathbf{B}_k^\Delta) \mathbf{u}_k + \mathbf{f} + \mathbf{f}_k^\Delta, \\ \mathbf{y}_k &= (\mathbf{C} + \mathbf{C}_k^\Delta) \mathbf{x}_k + \boldsymbol{\eta}_k \end{aligned} \quad (1)$$

where $\mathbf{x}_k, \mathbf{u}_k, \mathbf{y}_k$ are the state, input and output at time k , $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{f}$ are the system matrices; and $\mathbf{A}_k^\Delta, \mathbf{B}_k^\Delta, \mathbf{C}_k^\Delta, \mathbf{f}_k^\Delta$ are the weighted uncertain variable matrices affecting the system at time k . In particular, for each $\boldsymbol{\tau}_k^\Delta$, $\boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{f}\}$, we assume:

$$\boldsymbol{\tau}_k^\Delta = \boldsymbol{\tau}^N \odot \Delta_k^\tau \quad (2)$$

for some normalization matrix $\boldsymbol{\tau}^N$ and uncertainty matrix Δ_k^τ whose entries satisfy

$$-1 \leq \Delta_k^{\tau^{m,l}} \leq 1, \quad (3)$$

where $\Delta_k^{\tau^{m,l}}$ corresponds to the uncertainty variable associated with the term on m^{th} row and l^{th} column of parameter τ .

For simplicity of notation, we collect the entries of Δ_k^τ for all $\boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{f}\}$, in a vector $\Delta_k \in \Omega \subset \mathbb{R}^{n_\Delta}$, where n_Δ is the total number of uncertain parameters. It follows from the assumption (2) that the set Ω of admissible uncertainties is the unit infinity-norm ball in \mathbb{R}^{n_Δ} .

Now, we define SWA models, the main modeling framework used in this paper.

Definition 2 (*Switched Affine Model*) A switched affine (SWA) model is defined by:

$$\mathcal{G} = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i^\Delta\}_{i=1}^s), \quad (4)$$

where $\mathcal{X} \subset \mathbb{R}^n$ is the set of states, $\mathcal{E} \subset \mathbb{R}^{n_y}$ is the set of measurement noise values, $\mathcal{U} \subset \mathbb{R}^{n_u}$ is the set of inputs. The collection $\{G_i^\Delta\}_{i=1}^s$ is the set of s affine models. The evolution of \mathcal{G} is governed by:

$$\begin{aligned} \mathbf{x}_{k+1} &= (\mathbf{A}_{\sigma_k} + \mathbf{A}_{\sigma_k, k}^\Delta) \mathbf{x}_k + (\mathbf{B}_{\sigma_k} + \mathbf{B}_{\sigma_k, k}^\Delta) \mathbf{u}_k + \mathbf{f}_{\sigma_k} + \mathbf{f}_{\sigma_k, k}^\Delta, \\ \mathbf{y}_k &= (\mathbf{C}_{\sigma_k} + \mathbf{C}_{\sigma_k, k}^\Delta) \mathbf{x}_k + \boldsymbol{\eta}_k \end{aligned} \quad (5)$$

where $\sigma_k \in \{1, \dots, s\}$ indicates the active affine model (i.e. the mode) at time k , and $\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i, \mathbf{f}_i$ are the system matrices of the affine model G_i^Δ .

In what follows, we assume that the sets of admissible states, inputs and noise are hyper-rectangles defined as:

$$\mathcal{X} = \{\mathbf{x} \mid X_l \leq \mathbf{x} \leq X_u\}, \quad (6)$$

$$\mathcal{U} = \{\mathbf{u} \mid U_l \leq \mathbf{u} \leq U_u\}, \quad (7)$$

$$\mathcal{E} = \{\boldsymbol{\eta} \mid \epsilon_l \leq \boldsymbol{\eta} \leq \epsilon_u\}, \quad (8)$$

where subscript l and u denote lower and upper bounds, respectively. The bounds can be taken to be infinite.

Remark 1 Note that for the sake of simplicity in notation, we omit process noise, possible affine term in the output equation and the feed-forward term in the output equation in Def. 1. It is, however, straightforward to apply the proposed techniques when considering all the mentioned terms. Similarly, the uncertainty model can also be generalized to capture some correlations between entries.

3 Problem Statements

In this section, we present some preliminary definitions together with the two problems, solutions of which provide the basis of the proposed fault detection approach.

Definition 3 (*length- N behavior*) The length- N behavior associated with an SWA model \mathcal{G} is the set of all length- N input-output trajectories compatible with \mathcal{G} , given by the set

$$\begin{aligned} \mathcal{B}_{swa}^N(\mathcal{G}) &\doteq \left\{ \{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1} \mid \mathbf{u}_k \in \mathcal{U}, \exists \mathbf{x}_k \in \mathcal{X}, \boldsymbol{\eta}_k \in \mathcal{E}, \Delta_k \in \Omega, \sigma_k \in \mathbb{Z}_s^+, \right. \\ &\quad \left. (5) \text{ holds for all } k \in \mathbb{Z}_{N-1}^0 \right\}. \end{aligned}$$

With slight abuse of terminology, when N is clear from the context, we call $\mathcal{B}_{swa}^N(\mathcal{G})$ just the behavior of \mathcal{G} .

We also define behavior of an SWA model associated with a certain initial condition.

Definition 4 (*length- N state-initiated behavior*) The length- N behavior of an SWA model \mathcal{G} initiated at state \mathbf{x}^* is the set of all length- N input-output trajectories compatible with \mathcal{G} when the initial state is \mathbf{x}^* , given by the set

$$\mathcal{B}_{swa}^N(\mathcal{G}, \mathbf{x}^*) \doteq \{ \{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^{N-1} \mid \mathbf{u}_k \in \mathcal{U}, \exists \mathbf{x}_k \in \mathcal{X}, \mathbf{x}_0 = \mathbf{x}^*, \boldsymbol{\eta}_k \in \mathcal{E}, \Delta_k \in \Omega, \sigma_k \in \mathbb{Z}_s^+, (5) \text{ holds for all } k \in \mathbb{Z}_{N-1}^0 \}.$$

The consistency set associated with a behavior or state-initiated behavior of an SWA model is defined as follows.

Definition 5 (*consistency set*) Let $\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^N$ be an input-output trajectory over a time window $[0, N]$. The consistency set associated with $\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^N$ of an SWA model \mathcal{G} is defined as follows:

$$\mathcal{T}_{\mathcal{G}}(\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^N) \doteq \{ \mathbf{x}_{0:N}, \boldsymbol{\eta}_{0:N}, \Delta_{0:N}, \sigma_{0:N} \mid (5)-(8) \text{ hold for all } k \in \mathbb{Z}_N^0 \}, \quad (9)$$

$$\mathcal{T}_{\mathcal{G}}^*(\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^N, \mathbf{x}^*) \doteq \{ \mathbf{x}_{0:N}, \boldsymbol{\eta}_{0:N}, \Delta_{0:N}, \sigma_{0:N} \mid (5)-(8) \text{ hold for all } k \in \mathbb{Z}_N^0, \text{ and } \mathbf{x}_0 = \mathbf{x}^* \}, \quad (10)$$

where subscript $0:N$ indicates all the samples between times 0 and N .

In words, the consistency set is the set of state, noise, uncertainty and mode sequences that evolve through dynamics of the SWA model (5), and given the input sequence $\{ \mathbf{u}_k \}_{k=0}^N$ generate the output sequence $\{ \mathbf{y}_k \}_{k=0}^N$. Note that as the input signal is given, if the bounds on u_k are violated, it can be checked separately, therefore they are ignored in the definition of consistency sets.

The first problem we address in this paper is the model invalidation problem. Roughly speaking, given an input-output trajectory and an SWA model, the model invalidation problem is to determine whether or not the data is compatible with the model. This problem can be formally stated in terms of behaviors or consistency sets as follows.

Problem 1 Given $\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^{N-1}$, an input-output trajectory, and an SWA model \mathcal{G} , determine whether or not the input-output trajectory is contained in the behavior of \mathcal{G} . That is, determine whether or not the following is true

$$\{ \mathbf{u}_k, \mathbf{y}_k \}_{k=0}^{N-1} \in \mathcal{B}_{swa}^N(\mathcal{G}), \quad (11)$$

or equivalently $\mathcal{T}_{\mathcal{G}}(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1}) \neq \emptyset$.

For the well-posedness of the problems studied, we assume that $\mathcal{B}_{swa}^N(\mathcal{G}) \neq \emptyset$ for any positive integer N . This guarantees that the model is capable of generating some infinite trajectories; therefore, it does not invalidate itself if one waits long enough.

The abnormal trajectories for a system are those that cannot be generated by the model of a system.

Definition 6 (*abnormal trajectory*) An input-output trajectory $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1}$ is called abnormal for an SWA model \mathcal{G} if

$$\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1} \notin \mathcal{B}_{swa}^N(\mathcal{G}),$$

or equivalently $\mathcal{T}_{\mathcal{G}}(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^N) = \emptyset$.

With this definition, it is clear that a trajectory being abnormal is equivalent to the model being invalid for that trajectory. Therefore, a solution to Problem 1 can be readily used to detect abnormal trajectories or anomalies. Note that detecting abnormal trajectories does not require explicit models for the anomaly. Given that a cyber-physical system can fail (or be attacked) in infinitely many different ways, not needing to model these failure modes is advantageous. On the other hand, if one has an explicit model for a given fault, then this information can be used to develop more efficient fault detection schemes. In this paper, we represent faults also using SWA models.

Definition 7 (*fault*) A fault model for a system with an SWA model $\mathcal{G} = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i^\Delta\}_{i=1}^s)$ is another SWA model $\mathcal{G}^f = (\bar{\mathcal{X}}, \bar{\mathcal{E}}, \bar{\mathcal{U}}, \{\bar{G}_i^\Delta\}_{i=1}^{\bar{s}})$ with the same number of states, inputs and outputs.

In general, we expect faults to lead to abnormal trajectories. One can argue that a fault that does not lead to an abnormal trajectory, might not be that important from an operational standpoint. Moreover, it is not possible to detect a fault that has trajectories that are identical to the system trajectories. Next, we define T -detectability, which is a property of a given pair of SWA fault and system models, that measures how long it takes for a fault to lead to an abnormal trajectory.

Definition 8 (*T -detectability*) A fault model \mathcal{G}^f for a system model \mathcal{G} is called T -detectable¹ if $\mathcal{B}_{swa}^T(\mathcal{G}, \mathbf{x}^*) \cap \mathcal{B}_{swa}^T(\mathcal{G}^f, \mathbf{x}^*) = \emptyset$, for all $\mathbf{x}^* \in \mathcal{X}$, where T is a positive integer.

¹This notion is symmetric and therefore we also say \mathcal{G} and \mathcal{G}^f are T -detectable.

Note that, by definition, $\mathcal{B}_{swa}^N(\mathcal{G}, \mathbf{x}^*) = \emptyset$ if $\mathbf{x}^* \notin \mathcal{X}$ (the domain of \mathcal{G}). Clearly, if a fault is T -detectable for a system, then it is T^* -detectable for the same system for all $T^* \geq T$.

The second problem we are interested in is the characterization of the T -detectability property for a pair of fault and system models.

Problem 2 *Given two SWA models, \mathcal{G} and \mathcal{G}^f , and an integer T , determine whether the fault model \mathcal{G}^f is T -detectable for the system model \mathcal{G} , or not. That is, check if the set*

$$\mathcal{B}_{swa}^T(\mathcal{G}, \mathbf{x}^*) \cap \mathcal{B}_{swa}^T(\mathcal{G}^f, \mathbf{x}^*), \quad (12)$$

is empty for all $\mathbf{x}^ \in \mathcal{X}$ or not.*

As we show later in the paper, if a fault model is T -detectable for a system model, then a solution to the model invalidation problem can be used to efficiently detect faults without compromising the detection guarantees.

4 Fault Detection Scheme

In this section, we propose a fault detection scheme based on model invalidation. We initially present optimization based solutions to model invalidation and T -detectability problems. Additionally, we introduce T -weak detectability property that allows us to incorporate constraints on the mode sequences. Finally, we present an efficient fault detection scheme for T -(weak) detectable faults.

4.1 Model Invalidation

As discussed in Section 3, given a model of a system, detecting an abnormal trajectory is equivalent to model being invalid. Therefore, the solution to model invalidation problem, can be readily applied for detecting anomalies in a system from the input/output measurements. Next, we define a series of feasibility problems that are equivalent to Problem 1. That is, for a given input-output trajectory $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1}$, we encode the consistency set $\mathcal{T}_{\mathcal{G}}(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^N)$ with a feasibility problem. Since we have a hidden-mode switched system model, at each time, we want the data to satisfy the dynamic constraints for at least one mode. In order to capture this, consider the following mixed integer non-linear problem:

$$\begin{aligned} &\text{Find } \mathbf{x}_k, \boldsymbol{\eta}_k, \Delta_k, a_{i,k}, \forall k \in \mathbb{Z}_{N-1}^0, \forall i \in \mathbb{Z}_s^+ \\ &\text{s.t. } \forall k \in \mathbb{Z}_{N-1}^0, \forall i \in \mathbb{Z}_s^+ : \end{aligned} \quad (13)$$

$$\begin{aligned}
a_{i,k}\mathbf{x}_{k+1} &= a_{i,k}(\mathbf{A}_i + \mathbf{A}_{i,k}^\Delta)\mathbf{x}_k + a_{i,k}(\mathbf{B}_i + \mathbf{B}_{i,k}^\Delta)\mathbf{u}_k + a_{i,k}\mathbf{f}_i + a_{i,k}\mathbf{f}_{i,k}^\Delta, & (13a) \\
a_{i,k}\mathbf{y}_k &= a_{i,k}(\mathbf{C}_i + \mathbf{C}_{i,k}^\Delta)\mathbf{x}_k + a_{i,k}\boldsymbol{\eta}_k, & (13b) \\
a_{i,k}\boldsymbol{\tau}_{i,k}^\Delta &= a_{i,k}\boldsymbol{\tau}_i^N \odot \Delta_{i,k}^\tau, & (13c) \\
\sum_{i \in \mathbb{Z}_s^+} a_{i,k} &= 1, \quad a_{i,k} \in \{0, 1\}, & (13d) \\
X_l \leq \mathbf{x}_k \leq X_u, \quad \epsilon_l \leq \boldsymbol{\eta}_k \leq \epsilon_u, & & (13e) \\
|\Delta_{i,k}^{\tau^{m,l}}| \leq 1, \quad \boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{C}, \mathbf{B}, \mathbf{f}\}. & & (13f)
\end{aligned}$$

The constraints (13d) on the binary variables $a_{i,k}$ ensure that the data is compatible with at least one of the modes i at each time k . Next, we reformulate the problem (13) as the feasibility of a MILP problem, by leveraging ideas from convex hull formulation of mixed integer programming [44]. Note that in (13a)-(13c), there exist multi-affine terms in the form of products of state, uncertainty and binary variables. In what follows, we first apply a change of variables that renders these constraints linear. However, the constraints (13e)-(13f) are no longer linear in the new variables. Finally, we replace (13e)-(13f) with equivalent constraints that are linear in the new variables.

Let us start by introducing the following variables $\boldsymbol{\delta}_{i,k}^A \doteq a_{i,k}\mathbf{A}_{i,k}^\Delta\mathbf{x}_k$, $\boldsymbol{\delta}_{i,k}^C \doteq a_{i,k}\mathbf{C}_{i,k}^\Delta\mathbf{x}_k$. Here, $\boldsymbol{\delta}_{i,k}^\tau$ is a vector with m^{th} element equal to:

$$\boldsymbol{\delta}_{i,k}^{\tau^m} \doteq \sum_{l \in \mathbb{Z}_n^+} \underbrace{a_{i,k}[\boldsymbol{\tau}^N]^{m,l} \Delta_{i,k}^{\tau^{m,l}} \mathbf{x}_k^l}_{\doteq \boldsymbol{\delta}_{i,k}^{\tau^{m_l}}}, \quad \boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{C}\}. \quad (14)$$

The uncertain parameters in \mathbf{B} and \mathbf{f} only exist as the product of a binary variable and an uncertainty variable, hence we define:

$$\boldsymbol{\delta}_{i,k}^B \doteq a_{i,k}\mathbf{B}_{i,k}^\Delta, \quad \boldsymbol{\delta}_{i,k}^f \doteq a_{i,k}\mathbf{f}_{i,k}^\Delta. \quad (15)$$

Considering the change of variables in Table 1, we rewrite (13a) -(13d) as

Table 1: Change of variables for the model invalidation problem.

$$\begin{array}{ccc}
\mathbf{z}_{i,k} \doteq a_{i,k}\mathbf{x}_{k+1} & \mathbf{z}'_{i,k} \doteq a_{i,k}\mathbf{x}_k & \boldsymbol{\theta}_{i,k} \doteq a_{i,k}\boldsymbol{\eta}_k
\end{array}$$

follows:

$$\mathbf{z}_{i,k} = \mathbf{A}_i\mathbf{z}'_{i,k} + \boldsymbol{\delta}_{i,k}^A + a_{i,k}(\mathbf{B}_i\mathbf{u}_k + \mathbf{f}_k) + \boldsymbol{\delta}_{i,k}^B\mathbf{u}_k + \boldsymbol{\delta}_{i,k}^f, \quad (16a)$$

$$a_{i,k}\mathbf{y}_k = \mathbf{C}_i\mathbf{z}_{i,k} + \boldsymbol{\delta}_{i,k}^C + \boldsymbol{\theta}_{i,k}, \quad (16b)$$

$$\sum_{i \in \mathbb{Z}_s^+} a_{i,k} = 1, \quad a_{i,k} \in \{0, 1\}, \quad (16c)$$

$$\sum_{i \in \mathbb{Z}_s^+} \mathbf{z}_{i,k} = \sum_{i \in \mathbb{Z}_s^+} \mathbf{z}'_{i,k+1}. \quad (16d)$$

Constraints (16) are linear in the new variables. Bound constraints for the new variables can be obtained using conic equivalences [45]. In particular, constraints (13c) and (13e)-(13f) are equivalent to the following constraints in the new variables:

$$a_{i,k}X_l \leq \mathbf{z}'_{i,k} \leq a_{i,k}X_u, \quad a_{i,k}\epsilon_l \leq \boldsymbol{\theta}_{i,k} \leq a_{i,k}\epsilon_u, \quad (17a)$$

$$\forall m, l \in \mathbb{Z}_n^+, \quad \forall o \in \mathbb{Z}_{n_y}^+, \quad \forall p \in \mathbb{Z}_{n_u}^+ :$$

$$|\boldsymbol{\delta}_{i,k}^{B^{m,p}}| \leq a_{i,k} |[\mathbf{B}_i^N]^{m,p}|, \quad |\boldsymbol{\delta}_{i,k}^{f^m}| \leq a_{i,k} |[\mathbf{f}_i^N]^m|, \quad (17b)$$

$$|\boldsymbol{\delta}_{i,k}^{A^{ml}}| \leq |[\mathbf{A}_i^N]^{m,l}| |\mathbf{z}_{i,k}^l|, \quad |\boldsymbol{\delta}_{i,k}^{C^{ol}}| \leq |[\mathbf{C}_i^N]^{o,l}| |\mathbf{z}_{i,k}^l|. \quad (17c)$$

Proposition 1 *Given an SWA model \mathcal{G} and an input-output trajectory $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1}$, the model is invalidated if and only if the following problem is infeasible*

$$\begin{aligned} & \text{Find } \mathbf{z}_{i,k}, \mathbf{z}'_{i,k}, \boldsymbol{\theta}_{i,k}, \boldsymbol{\delta}_{i,k}^A, \boldsymbol{\delta}_{i,k}^B, \boldsymbol{\delta}_{i,k}^C, \boldsymbol{\delta}_{i,k}^f, a_{i,k}, \forall i \in \mathbb{Z}_s^+ \forall k \in \mathbb{Z}_{N-1}^0 \quad (\text{P}_{MI}) \\ & \text{s.t. } \{(16), (17)\}, \quad \forall k \in \mathbb{Z}_{N-1}^0, \quad \forall i \in \mathbb{Z}_s^+. \end{aligned}$$

Proof: Since (13) encodes the consistency set, invalidation is clearly equivalent to its infeasibility. Therefore it is enough to show the equivalence of (13) and (P_{MI}). Given a feasible point of (13), a feasible point of (P_{MI}) can be constructed by applying the change of variables introduced above. For the other direction, let $\mathbf{z}_{i,k}^*, \mathbf{z}'_{i,k}, \boldsymbol{\theta}_{i,k}^*, a_{i,k}^*, \boldsymbol{\delta}_{i,k}^{A*}, \boldsymbol{\delta}_{i,k}^{B*}, \boldsymbol{\delta}_{i,k}^{C*}, \boldsymbol{\delta}_{i,k}^{f*}$ be a feasible point of (P_{MI}). We will construct a feasible $\mathbf{x}_k^\circ, \boldsymbol{\eta}_k^\circ, a_{i,k}^\circ, \Delta_{i,k}^{A^\circ}, \Delta_{i,k}^{B^\circ}, \Delta_{i,k}^{C^\circ}, \Delta_{i,k}^{f^\circ}$ for (13). Take $a_{i,k}^\circ = a_{i,k}^*$. Since the constraints (13d) and (16c) are identical, $a_{i,k}^\circ$ satisfies (13d). Moreover, at each time k , there is a unique mode i_k^* such that $a_{i_k^*,k}^\circ = 1$ and $a_{i,k}^\circ = 0, \forall i \neq i_k^*$.

First consider the constraints in (P_{MI}) corresponding to the mode sequence $\{i_k^*\}_{k=0}^{N-1}$. From (15), take $\mathbf{B}^{\Delta_{i_k^*,k}^\circ} = \boldsymbol{\delta}_{i_k^*,k}^{B*}$ and $\mathbf{f}^{\Delta_{i_k^*,k}^\circ} = \boldsymbol{\delta}_{i_k^*,k}^{f*}$. From Table 1, we have (i) $\boldsymbol{\eta}_k^\circ = \boldsymbol{\theta}_{i_k^*,k}^*$; and (ii) $\mathbf{z}_{i,k}^* = \mathbf{z}'_{i,k}^* = 0$ for $i \neq i_k^*$, which when combined with (16d) ensures that $\mathbf{z}_{i_k^*,k}^* = \mathbf{z}'_{i_k^*,k+1}^*$ and helps us uniquely construct $\mathbf{x}_k^\circ = \mathbf{z}_{i_k^*,k}^*$. Given $\mathbf{x}_k^\circ, a_{i,k}^\circ$ and $\boldsymbol{\delta}_{i,k}^{\tau*}$, and assuming $\boldsymbol{\tau}^{N^\circ}$ is nonzero, constraint (14) helps us construct $\Delta_{i_k^*,k}^{\tau^\circ}$ for $\tau \in \{\mathbf{A}, \mathbf{C}\}$. Note that if $\boldsymbol{\tau}^{N^\circ} = 0$, then $\Delta_{i_k^*,k}^{\tau^\circ}$ is also zero. By construction, these variables satisfy the constraints (13a)-(13c) for the active mode i_k^* at time k , and (13e)-(13f) for all k . At each time k for all the inactive modes $i \neq i_k^*$, we have $a_{i,k}^\circ = 0$, hence one can arbitrarily pick the uncertainty variables within their constraint set (13f) and the constraints (13a)-(13c) are trivially satisfied for all (i, k) in the set $\{(i, k) \mid i \in \mathbb{Z}_s^+, k \in \mathbb{Z}_{N-1}^0, (i, k) \neq (i_k^*, k)\}$, which concludes the proof. \square

We refer to an instance of the problem (P_{MI}) for a given SWA model \mathcal{G} and an input-output trajectory $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1}$ as $\text{Feas}_{\mathcal{G}}(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1})$ in the remainder of the paper. Both the objective function and the constraints of the problem (P_{MI}) are linear except for the constraints (17b)-(17c). However, it is straightforward to transform this constraint into a set of linear constraints as shown in Appendix A. After this transformation the problem (P_{MI}) becomes a MILP problem in feasibility form.

4.2 T -Detectability

In general, for detecting an arbitrary, unknown fault (or anomaly), one needs to use all the data that is available, which leads to larger and larger problems as time passes. On the other hand, as eluded to earlier, the availability of a fault model can be exploited to develop more efficient fault detection schemes while preserving the detection guarantees. We are interested in T -detectability of the pair of system and fault models as a key property that enables us to apply the model invalidation approach for fault detection in a receding horizon manner, rather than applying it on the entire time horizon, while preserving all the guarantees for detection. Fig. 1 illustrates the motivation of introducing the concept of T -detectability. In this section, we develop a MILP characterization of T -detectability property that can be verified off-line. The following proposition formalizes the fact that T -

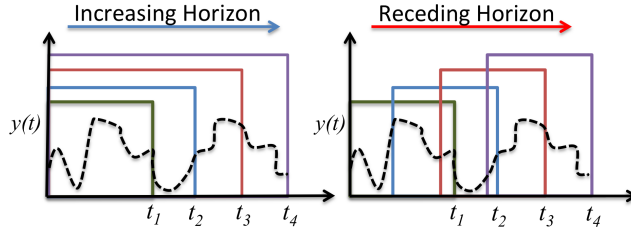


Figure 1: Left: without T -detectability, the model invalidation has to be applied on an increasing-size data. Right: with T -detectability, the model invalidation is applied in a receding horizon manner.

detectable faults can be detected with a receding horizon algorithm.

Proposition 2 *Given a T -detectable fault model \mathcal{G}^f for SWA model \mathcal{G} , it is possible to detect the existence of a persistent fault evolving through \mathcal{G}^f by checking, at each time k , if $\text{Feas}_{\mathcal{G}}(\{\mathbf{u}_j, \mathbf{y}_j\}_{j=k-T+1}^k)$ is feasible or not.*

Proof: Let a fault occur at time i^* ; that is, the input-output trajectory $\{\mathbf{u}_j, \mathbf{y}_j\}_{j \geq i^*}$ is generated by the fault model \mathcal{G}^f . Because \mathcal{G}^f is T -detectable, by Def. 8, there exists $k^* \leq i^* + T - 1$ such that $\{\mathbf{u}_j, \mathbf{y}_j\}_{j=i^*}^{k^*} \notin \mathcal{B}_{swa}^{k^*-i^*+1}(\mathcal{G})$. By Proposition 1, this is equivalent to the existence of $k^* \leq i^* + T - 1$ such that $\text{Feas}_{\mathcal{G}}(\{\mathbf{u}_j, \mathbf{y}_j\}_{j=i^*}^{k^*})$ is infeasible. Since $[i^*, k^*] \subseteq [k^* - T + 1, k^*]$, the infeasibility of $\text{Feas}_{\mathcal{G}}(\{\mathbf{u}_j, \mathbf{y}_j\}_{j=i^*}^{k^*})$ implies the infeasibility of $\text{Feas}_{\mathcal{G}}(\{\mathbf{u}_j, \mathbf{y}_j\}_{j=k^*-T+1}^{k^*})$. \square

Now that we have an efficient way to solve fault detection problem for a T -detectable fault, the next question is how to check, given an integer T , whether a fault is T -detectable for a particular system. In what follows, we give a necessary and sufficient condition under which a fault \mathcal{G}^f is T -detectable for a system \mathcal{G} that can be checked by checking the feasibility of MILP certificates. First we make the following assumptions:

Assumption 1 *Faults are permanent.*

Assumption 2 *At the initial time (i.e., $k = 0$), the system is healthy.*

Remark 2 *In Section 5, we show how to relax Assumption 1. As for Assumption 2, it enables us to use state-initiated behaviors in the detectability definition instead of the behaviors. It is reasonable to assume that when a system is deployed, it starts at a healthy condition and when a fault occurs the behaviors of the healthy system and faulty system start to deviate from the same initial condition. This assumption can easily be omitted, as is done in our earlier work [10], leading to a stronger (harder to satisfy) definition of T -detectability that involves behaviors instead of state-initiated behaviors.*

In terms of consistency sets, \mathcal{G}^f is T -detectable for \mathcal{G} if for any $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^T \in \mathcal{B}_{swa}^T(\mathcal{G}^f, \mathbf{x}^*)$ and $\mathbf{x}^* \in \mathcal{X}$, we have $\mathcal{T}_{\mathcal{G}}^*(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^T, \mathbf{x}^*) = \emptyset$, or equivalently for any $\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^T$ and \mathbf{x}^* , the following holds:

$$\mathcal{T}_{\mathcal{G}}^*(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^T, \mathbf{x}^*) \cap \mathcal{T}_{\mathcal{G}^f}^*(\{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^T, \mathbf{x}^*) = \emptyset. \quad (18)$$

Consider the following nonlinear feasibility problem:

$$\begin{aligned} &\text{Find } \mathbf{x}_k, \bar{\mathbf{x}}_k, \boldsymbol{\eta}_k, \bar{\boldsymbol{\eta}}_k, \mathbf{u}_k, \Delta_k, \bar{\Delta}_k, d_{i,j,k}, \forall k \in \mathbb{Z}_{T-1}^0, \forall i \in \mathbb{Z}_s^+, \forall j \in \mathbb{Z}_s^+ \\ &\text{s.t. } \forall k \in \mathbb{Z}_{T-1}^0, \forall i \in \mathbb{Z}_s^+, \forall j \in \mathbb{Z}_s^+ : \end{aligned} \quad (19)$$

$$\begin{aligned}
& d_{i,j,k}(\mathbf{x}_{k+1} - (\mathbf{A}_i + \mathbf{A}_{i,k}^\Delta)\mathbf{x}_k - (\mathbf{B}_i + \mathbf{B}_{i,k}^\Delta)\mathbf{u}_k - \mathbf{f}_i - \mathbf{f}_{i,k}^\Delta) = 0, & (19a) \\
& d_{i,j,k}(\bar{\mathbf{x}}_{k+1} - (\bar{\mathbf{A}}_j + \bar{\mathbf{A}}_{j,k}^\Delta)\bar{\mathbf{x}}_k - (\bar{\mathbf{B}}_j + \bar{\mathbf{B}}_{j,k}^\Delta)\mathbf{u}_k - \bar{\mathbf{f}}_j - \bar{\mathbf{f}}_{j,k}^\Delta) = 0, & (19b) \\
& d_{i,j,k}((\mathbf{C}_i + \mathbf{C}_{i,k}^\Delta)\mathbf{x}_k + \boldsymbol{\eta}_k - (\bar{\mathbf{C}}_j + \bar{\mathbf{C}}_{j,k}^\Delta)\bar{\mathbf{x}}_k - \bar{\boldsymbol{\eta}}_k) = 0, & (19c) \\
& \mathbf{x}_0 = \bar{\mathbf{x}}_0, & (19d) \\
& \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_{\bar{s}}^+} d_{i,j,k} = 1, \quad d_{i,j,k} \in \{0, 1\}, & (19e) \\
& X_l \leq \mathbf{x}_k \leq X_u, \quad \bar{X}_l \leq \bar{\mathbf{x}}_k \leq \bar{X}_u, & (19f) \\
& \epsilon_l \leq \boldsymbol{\eta}_k \leq \epsilon_u, \quad \bar{\epsilon}_l \leq \bar{\boldsymbol{\eta}}_k \leq \bar{\epsilon}_u, \quad U_l \leq \mathbf{u}_k \leq U_u, & (19g) \\
& \boldsymbol{\tau}_{i,k}^\Delta = \boldsymbol{\tau}_i^N \odot \Delta_{i,k}^\tau, \quad |\Delta_{i,k}^{\tau^{m,l}}| \leq 1, \boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{C}, \mathbf{B}, \mathbf{f}\}, & (19h) \\
& \bar{\boldsymbol{\tau}}_{j,k}^\Delta = \bar{\boldsymbol{\tau}}_j^N \odot \bar{\Delta}_{j,k}^\tau, \quad |\bar{\Delta}_{j,k}^{\tau^{m,l}}| \leq 1, \bar{\boldsymbol{\tau}} \in \{\bar{\mathbf{A}}, \bar{\mathbf{C}}, \bar{\mathbf{B}}, \bar{\mathbf{f}}\}. & (19i)
\end{aligned}$$

Clearly, the infeasibility of (19) is equivalent to (18) being true, therefore, the T -detectability of \mathcal{G}^f for \mathcal{G} . This is stated formally below.

Proposition 3 *The fault model \mathcal{G}^f is T -detectable for the system \mathcal{G} if and only if problem (19) is infeasible.*

Proof: The binary variables $d_{i,j,k}$ in (19) guarantee that at each time k , there is a mode j of the fault model \mathcal{G}^f and a mode i of the system model \mathcal{G} such that the outputs of the two models match. Since both models' initial states are enforced to be equal, the rest follows from the definitions of the state-initiated behaviors and consistency sets. \square

Following steps similar to those in the model invalidation subsection, we derive an MILP problem equivalent to Problem (19). In order to eliminate the multi-affine terms in (19), we consider the following change of variables: $\boldsymbol{\delta}_{i,j,k}^A \doteq d_{i,j,k} \mathbf{A}_{i,k}^\Delta \mathbf{x}_k$, $\boldsymbol{\delta}_{i,j,k}^C \doteq d_{i,j,k} \mathbf{C}_{i,k}^\Delta \mathbf{x}_k$, $\bar{\boldsymbol{\delta}}_{i,j,k}^A \doteq d_{i,j,k} \bar{\mathbf{A}}_{j,k}^\Delta \bar{\mathbf{x}}_k$, and $\bar{\boldsymbol{\delta}}_{i,j,k}^C \doteq d_{i,j,k} \bar{\mathbf{C}}_{j,k}^\Delta \bar{\mathbf{x}}_k$. Here, $\boldsymbol{\delta}_{i,j,k}^\tau$ and $\bar{\boldsymbol{\delta}}_{i,j,k}^\tau$ are vectors with m^{th} elements equal to:

$$\begin{aligned}
\boldsymbol{\delta}_{i,k}^{\tau^m} & \doteq \sum_{l \in \mathbb{Z}_n^+} \underbrace{d_{i,j,k} [\boldsymbol{\tau}^N]^{m,l} \Delta_{i,k}^{\tau^{m,l}} \mathbf{x}_k^l}_{\doteq \boldsymbol{\delta}_{i,j,k}^{\tau^{m,l}}}, \quad \boldsymbol{\tau} \in \{\mathbf{A}, \mathbf{C}\}, \\
\bar{\boldsymbol{\delta}}_{i,k}^{\tau^m} & \doteq \sum_{l \in \mathbb{Z}_n^+} \underbrace{d_{i,j,k} [\bar{\boldsymbol{\tau}}^N]^{m,l} \bar{\Delta}_{i,k}^{\tau^{m,l}} \bar{\mathbf{x}}_k^l}_{\doteq \bar{\boldsymbol{\delta}}_{i,j,k}^{\tau^{m,l}}}, \quad \boldsymbol{\tau} \in \{\bar{\mathbf{A}}, \bar{\mathbf{C}}\}.
\end{aligned} \tag{20}$$

For uncertainty in parameters \mathbf{B} a similar change of variables leads to: $\boldsymbol{\delta}_{i,j,k}^B \doteq d_{i,j,k} \mathbf{B}_{i,k}^\Delta \mathbf{u}_k$ and $\bar{\boldsymbol{\delta}}_{i,j,k}^B \doteq d_{i,j,k} \bar{\mathbf{B}}_{j,k}^\Delta \mathbf{u}_k$. Note that in contrast to model invalidation problem, inputs are variables for the T -detectability problem.

The m^{th} row of the vectors $\delta_{i,j,k}^B$ is:

$$\delta_{i,j,k}^{\mathbf{B}^m} \doteq \sum_{p \in \mathbb{Z}_{nu}^+} \underbrace{d_{i,j,k} [\mathbf{B}^N]^{m,p} \Delta_{i,k}^{\mathbf{B}^{m,p}} \mathbf{u}_k^p}_{\doteq \delta_{i,j,k}^{\mathbf{B}^{mp}}}. \quad (21)$$

Corresponding variables $\bar{\delta}_{i,j,k}^B$ and $\bar{\delta}_{i,j,k}^{\mathbf{B}^{mp}}$ for the fault model are defined similarly. Finally, the uncertainty associated with the offset term \mathbf{f} only multiplies to the binary variable, hence we define: $\delta_{i,j,k}^f \doteq d_{i,j,k} \mathbf{f}_{i,k}^\Delta$, $\bar{\delta}_{i,j,k}^f \doteq d_{i,j,k} \bar{\mathbf{f}}_{j,k}^\Delta$. The remaining change of variables are given in Table 2.

Table 2: The list of change of variables for T -detectability Problem

$$\begin{aligned} \mathbf{z}_{i,j,k} &\doteq d_{i,j,k} \mathbf{x}_{k+1} & \mathbf{z}'_{i,j,k} &\doteq d_{i,j,k} \mathbf{x}_k & \bar{\mathbf{z}}_{i,j,k} &\doteq d_{i,j,k} \bar{\mathbf{x}}_{k+1} & \mathbf{t}_{i,j,k} &\doteq d_{i,j,k} \mathbf{u}_k \\ \bar{\mathbf{z}}'_{i,j,k} &\doteq d_{i,j,k} \bar{\mathbf{x}}_k & \boldsymbol{\theta}_{i,j,k} &\doteq d_{i,j,k} \boldsymbol{\eta}_k & \bar{\boldsymbol{\theta}}_{i,j,k} &\doteq d_{i,j,k} \bar{\boldsymbol{\eta}}_k \end{aligned}$$

By plugging in the new variables we can equivalently state the constraints in (19) by the following MILP constraints:

$$\begin{aligned} \mathbf{z}_{i,j,k} &= \mathbf{A}_i \mathbf{z}'_{i,j,k} + \delta_{i,j,k}^A + \mathbf{B}_i \mathbf{t}_{i,j,k} + \delta_{i,j,k}^B + d_{i,j,k} \mathbf{f}_i + \delta_{i,j,k}^f, \\ \bar{\mathbf{z}}_{i,j,k} &= \bar{\mathbf{A}}_j \bar{\mathbf{z}}'_{i,j,k} + \bar{\delta}_{i,j,k}^A + \bar{\mathbf{B}}_j \mathbf{t}_{i,j,k} + \bar{\delta}_{i,j,k}^B + d_{i,j,k} \bar{\mathbf{f}}_j + \bar{\delta}_{i,j,k}^f, \\ \mathbf{C}_i \mathbf{z}'_{i,j,k} + \delta_{i,j,k}^C + \boldsymbol{\theta}_{i,j,k} &= \bar{\mathbf{C}}_j \bar{\mathbf{z}}'_{i,j,k} + \bar{\delta}_{i,j,k}^C + \bar{\boldsymbol{\theta}}_{i,j,k}, \\ \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} d_{i,j,k} &= 1, \quad d_{i,j,k} \in \{0, 1\}, \\ \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \mathbf{z}_{i,j,k} &= \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \mathbf{z}'_{i,j,k+1}, \\ \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \bar{\mathbf{z}}_{i,j,k} &= \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \bar{\mathbf{z}}'_{i,j,k+1}, \\ \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \mathbf{z}'_{i,j,0} &= \sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathbb{Z}_s^+} \bar{\mathbf{z}}'_{i,j,0}. \end{aligned} \quad (22)$$

Additionally, we transform the admissible set constraints, which are not linear in the new variables, to equivalent MILP certificates. The transformed

equivalent admissible set constraints are as follows:

$$\begin{aligned}
d_{i,j,k}X_l &\leq \mathbf{z}_{i,j,k} \leq d_{i,j,k}X_u, \quad d_{i,j,k}X_l \leq \mathbf{z}'_{i,j,k} \leq d_{i,j,k}X_u, \\
d_{i,j,k}\bar{X}_l &\leq \bar{\mathbf{z}}_{i,j,k} \leq d_{i,j,k}\bar{X}_u, \quad d_{i,j,k}\bar{X}_l \leq \bar{\mathbf{z}}'_{i,j,k} \leq d_{i,j,k}\bar{X}_u, \\
d_{i,j,k}\epsilon_l &\leq \boldsymbol{\theta}_{i,j,k} \leq d_{i,j,k}\epsilon_u, \quad d_{i,j,k}\bar{\epsilon}_l \leq \bar{\boldsymbol{\theta}}_{i,j,k} \leq d_{i,j,k}\bar{\epsilon}_u, \\
d_{i,j,k}U_l &\leq \mathbf{t}_{i,j,k} \leq d_{i,j,k}U_u, \\
\forall m, l \in \mathbb{Z}_n^+, \quad \forall o \in \mathbb{Z}_{n_y}^+, \quad \forall p \in \mathbb{Z}_{n_u}^+ : & \\
|\boldsymbol{\delta}_{i,j,k}^{A^{m_l}}| &\leq |[\mathbf{A}_i^N]^{m,l}| |\mathbf{z}_{i,j,k}^l|, \quad |\bar{\boldsymbol{\delta}}_{i,j,k}^{A^{m_l}}| \leq |[\bar{\mathbf{A}}_j^N]^{m,l}| |\bar{\mathbf{z}}_{i,j,k}^l| \\
|\boldsymbol{\delta}_{i,j,k}^{B^{mp}}| &\leq |[\mathbf{B}_i^N]^{m,p}| |\mathbf{t}_{i,j,k}^p|, \quad |\bar{\boldsymbol{\delta}}_{i,j,k}^{B^{mp}}| \leq |[\bar{\mathbf{B}}_j^N]^{m,p}| |\bar{\mathbf{t}}_{i,j,k}^p|, \\
|\boldsymbol{\delta}_{i,j,k}^{C^{ol}}| &\leq |[\mathbf{C}_i^N]^{o,l}| |\mathbf{z}_{i,j,k}^l|, \quad |\bar{\boldsymbol{\delta}}_{i,j,k}^{C^{ol}}| \leq |[\bar{\mathbf{C}}_j^N]^{o,l}| |\bar{\mathbf{z}}_{i,j,k}^l|, \\
|\boldsymbol{\delta}_{i,j,k}^{fm}| &\leq d_{i,j,k}, \quad |\bar{\boldsymbol{\delta}}_{i,j,k}^{fm}| \leq d_{i,j,k}.
\end{aligned} \tag{23}$$

These constraints ensure that all the variables corresponding to mode pairs (i, j) that are not active are zero, and for the active pair of modes, each variable is within its admissible bounds. Again the absolute value constraints can be converted to equivalent linear constraints as illustrated in Appendix A.

Theorem 1 *The fault model \mathcal{G}^f is T -detectable for the system \mathcal{G} if and only if (P_T) is infeasible.*

$$\begin{aligned}
&\text{Find } \mathbf{z}_{i,j,k}, \mathbf{z}'_{i,j,k}, \bar{\mathbf{z}}_{i,j,k}, \bar{\mathbf{z}}'_{i,j,k}, \boldsymbol{\theta}_{i,j,k}, \bar{\boldsymbol{\theta}}_{i,j,k}, \mathbf{t}_{i,j,k}, d_{i,j,k}\boldsymbol{\delta}_{i,j,k}^A, \bar{d}_{i,j,k}\bar{\boldsymbol{\delta}}_{i,j,k}^A, \boldsymbol{\delta}_{i,j,k}^B, \bar{\boldsymbol{\delta}}_{i,j,k}^B, \\
&\boldsymbol{\delta}_{i,j,k}^C, \bar{\boldsymbol{\delta}}_{i,j,k}^C, \boldsymbol{\delta}_{i,j,k}^f, \bar{\boldsymbol{\delta}}_{i,j,k}^f \\
&s.t. \{(22) - (23)\} \forall k \in \mathbb{Z}_{T-1}^0, \quad \forall i \in \mathbb{Z}_s^+, \quad \forall j \in \mathbb{Z}_s^+.
\end{aligned} \tag{P_T}$$

The proof is similar to that of Proposition 1 and omitted for space constraints.

Theorem 1 provides a necessary and sufficient condition for verifying, for a given pair of fault and system models and a given integer T , the T -detectability of the fault for the particular system. In general, to compute the minimum such T , one needs to start with $T = 1$ and solve a sequence of problems of the form (P_T) while incrementing T until infeasibility is achieved. Note that a finite T does not necessarily exist for all types of faults.

Next, we analyze T -detectability in a simplified setting with two autonomous affine models G, \bar{G} described by (1) with no uncertainty or noise.

Define observability matrix and projected affine terms for T measurements as follows:

$$\mathcal{O}_T(G) \doteq \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{T-1} \end{bmatrix}, \mathcal{F}_T(G) \doteq \begin{bmatrix} 0 \\ \mathbf{Cf} \\ \vdots \\ \sum_{j \in \mathbb{Z}_{T-2}^0} \mathbf{CA}^j \mathbf{f} \end{bmatrix}. \quad (24)$$

It is easy to see that, in this setting, a fault \bar{G} is not T -detectable for a system G for a given T if and only if there exists an initial condition \mathbf{x}_0 such that the outputs of two models are equal, that is,

$$\mathcal{O}_T(G)\mathbf{x}_0 + \mathcal{F}_T(G) = \mathcal{O}_T(\bar{G})\mathbf{x}_0 + \mathcal{F}_T(\bar{G}). \quad (25)$$

Our MILP formulation (P_T) of T -detectability essentially generalizes the condition (25) as in the case of switching, noise and uncertainty, the geometry of the consistency set cannot be represented by simple hyperplane conditions like this. A rearrangement of this condition gives us a necessary and sufficient condition on the existence of a finite T .

Proposition 4 *Let $\mathcal{O}_T(G, \bar{G}) = [\mathcal{O}_T(G) - \mathcal{O}_T(\bar{G})]$. Two autonomous affine models G, \bar{G} are **not** T -detectable for any finite T if and only if there exists an initial condition \mathbf{x}_0 such that*

$$\begin{bmatrix} \mathbf{C} - \bar{\mathbf{C}} \end{bmatrix} \mathbf{x}_0 = \mathbf{0} \quad (26)$$

$$\mathcal{O}_T(G, \bar{G}) \left(\begin{bmatrix} \mathbf{f} \\ \bar{\mathbf{f}} \end{bmatrix} - \begin{bmatrix} \mathbf{A} - I \\ \bar{\mathbf{A}} - I \end{bmatrix} \mathbf{x}_0 \right) = \mathbf{0}, \quad (27)$$

for $T = 2n$.

Proof: By manipulating (25), it is possible to arrive at the equivalent conditions (26) and (27). Hence, if no \mathbf{x}_0 satisfying (26) and (27) exists, there is a finite T ($T = 2n$). Now assume an \mathbf{x}_0 exists. Note that $\mathcal{O}_T(G, \bar{G})$ is the observability matrix of a system that is constructed by concatenating the states of the fault and system models, and taking the difference of their outputs as the output. Since $\mathcal{O}_T(G, \bar{G})$ is an observability matrix, its nullspace

remains the same beyond $T = 2n$. Therefore, the same \mathbf{x}_0 satisfies (27) for any finite $T > 2n$, from which we conclude that no finite T exists. \square

The conditions given in Proposition 4 resemble the mode discernibility and mode observability conditions for switched systems [34,35], with the difference that we consider affine models and enforce initial states to coincide. These conditions can also be used to arrive at simple sufficient conditions on non-existence of a finite T in the general setting. In particular, if there exist fixed values of input, noise, uncertainty and mode for two SWA models \mathcal{G} , \mathcal{G}^f such that (26) and (27) hold for the corresponding autonomous affine models for some initial condition, then \mathcal{G} and \mathcal{G}^f are not T -detectable for any finite T .

4.3 Weak Detectability

T -detectability, although useful to reduce the complexity of the proposed fault detection scheme, can be rather strong in the sense that it is not hard to encounter faults in real applications that are not T -detectable for any finite T . In this section, inspired by the indicators in discrete-event systems diagnosis problems [46], we propose the notion of weak detectability that incorporates language constraints on the hidden mode in the fault model. Since these indicators further restrict the behavior of the fault, it enlarges the class of faults that can be detected within a finite horizon.

Let us motivate weak detectability and indicators with an example. Consider a building radiant system with two modes that represent a controlled valve being open (mode 1) or closed (mode 2). When the control logic is not modeled, this model allows arbitrary switching between the two modes. Now consider a problem with the valve, which prevents it from being totally closed. Therefore the fault model consists of two modes: open and half-open. This fault model is not T -detectable because if the valve is never commanded to close (i.e., system always operates in mode 1), it is not possible to differentiate healthy system from the faulty one. On the other hand, this fault only affects the operation of the system when the valve is commanded close so it will not be of interest to detect it if the valve is never required to be closed. That is, the fault is relevant only if mode 2 becomes active. In order to incorporate such information in the fault models, we introduce indicators that shrink the behavior set of the SWA fault models by restricting their allowable mode sequences. Let us now formally define what we mean by an indicator.

Definition 9 (*indicator*) Given a fault model \mathcal{G}^f with s modes, let $(\mathbb{Z}_s^+)^W$

be the set of all length- W mode sequences. An indicator \mathcal{I} is a subset of $(\mathbb{Z}_s^+)^W$ for some W . The fault-indicator model $\mathcal{G}_{\mathcal{I}}^f$ is the fault model \mathcal{G}^f whose mode sequences are restricted to \mathcal{I} on the first W time steps after the fault occurs.

Indicators can be compactly defined using bounded regular languages or fixed-length prefixes of some linear temporal logic formula [47]. We also introduce a special class of indicators that we represent by tuples

$$\mathcal{I}_t \doteq (\mathcal{S}, W, m, O), O \in \{>, =, <\},$$

where \mathcal{I}_t consists of mode sequences of length W that contain modes from the set $\mathcal{S} \subseteq \mathbb{Z}_s^+$ more than (O is $>$), exactly (O is $=$), less than (O is $<$) m times. We have found this representation to be convenient in various applications we considered.

The behaviors of fault-indicator models are defined similarly.

Definition 10 (*state-initiated fault-indicator behavior*) *The length- N state-initiated behavior of the fault-indicator model $\mathcal{G}_{\mathcal{I}}^f$ from the initial state $\mathbf{x}^* \in \mathcal{X}$ is defined as:*

$$\begin{aligned} \mathcal{B}_{FI}^N(\mathcal{G}_{\mathcal{I}}^f, \mathbf{x}^*) &\doteq \{ \{\mathbf{u}_k, \mathbf{y}_k\}_{k=0}^{N-1} \mid \mathbf{u}_k \in \bar{\mathcal{U}} \text{ and } \exists \mathbf{x}_k \in \bar{\mathcal{X}}, \mathbf{x}_0 = \mathbf{x}^*, \boldsymbol{\eta}_k \in \bar{\mathcal{E}}, \\ &\Delta_k \in \bar{\Omega}, \{\sigma_k\}_{k=0}^{W-1} \in \mathcal{I} \text{ s.t. (5) holds for fault model parameters} \}. \end{aligned}$$

In Def. 10, we assume that $N \geq W$. It is possible to relax this assumption by considering length- N prefixes of sequences in \mathcal{I} . Now we are ready to define T -weak detectability.

Definition 11 (*T -weak detectability*) *The fault model \mathcal{G}^f is T -weak detectable with indicator \mathcal{I} for the system model \mathcal{G} , if the fault-indicator model $\mathcal{G}_{\mathcal{I}}^f$ is T -detectable for \mathcal{G} , that is, the following holds:*

$$\mathcal{B}_{swa}^T(\mathcal{G}, \mathbf{x}^*) \cap \mathcal{B}_{FI}^T(\mathcal{G}_{\mathcal{I}}^f, \mathbf{x}^*) = \emptyset. \quad (28)$$

Checking T -weak detectability also reduces to an equivalent MILP problem. Recall that the binary variable $d_{i,j,k}$ in (P_T) indicates if mode i of the system and mode j of the fault are active at time k , hence $\sum_{i \in \mathbb{Z}_s^+} d_{i,j,k} = 1$ indicates that mode j of the fault is active at time k . Then, if the mode sequence of the fault is $w_m \doteq \{\sigma_0^m \dots \sigma_{W-1}^m\} \in \mathcal{I}$, the following constraint needs to be enforced:

$$\sum_{i \in \mathbb{Z}_s^+} (d_{i,\sigma_0^m,0} + d_{i,\sigma_1^m,1} + \dots + d_{i,\sigma_{W-1}^m,W-1}) = W. \quad (29)$$

Since the mode sequence of the fault should follow at least one $w_m \in \mathcal{I}$, we have

$$b_m (\sum_{i \in \mathbb{Z}_s^+} (d_{i, \sigma_0^m, 0} + \dots + d_{i, \sigma_{W-1}^m, W-1}) - W) = 0 \quad (30)$$

with $\sum_{m \in \mathbb{Z}_\ell^+} b_m \geq 1$, where b_m are new binary variables and $\ell \doteq |\mathcal{I}|$. Finally, the change of variables, $d_{i, \sigma_k^m, k}^m \doteq b_m d_{i, \sigma_k^m, k}$, and extra constraints, $d_{i, \sigma_k^m, k}^m \leq b_m$, converts (30) to the equivalent MILP constraints :

$$\sum_{i \in \mathbb{Z}_s^+} d_{i, \sigma_k^m, k}^m - b_m W = 0, \quad d_{i, \sigma_k^m, k}^m \leq b_m, \quad \sum_{m \in \mathbb{Z}_\ell^+} b_m \geq 1. \quad (31)$$

In general, structured indicators can be expressed by much less number of MILP constraints. For instance the indicators of form $\mathcal{I}_t = (\mathcal{S}, W, m, >)$ can be expressed by a single MILP constraint:

$$\sum_{i \in \mathbb{Z}_s^+} \sum_{j \in \mathcal{S}} \sum_{k \in \mathbb{Z}_W^+} d_{i, j, k} \geq m. \quad (32)$$

4.4 Overall Fault Detection Scheme

The overall fault detection scheme consists of an offline step and an online step. Given a system and a fault model (possibly with indicators), we first analyze T -(weak) detectability running Algorithm 1 until it returns a T or maximum number of iterations are reached.

Algorithm 1 Calculating T

Input: $T_0, \mathcal{G}, \mathcal{G}_T^f$

1. Initialize $T = T_0$.
2. Set \mathcal{I}' to length- T prefixes of \mathcal{I}
3. Solve feasibility problem (P_T) with constraints (31) for \mathcal{I}'
4. If (P_T) is feasible:
 - $T = T + 1$ and go to step 2

Else

- Return T
-

Assuming that Algorithm 1 returns a finite T , at run-time, we solve the problem (P_{MI}) using data from the last T steps at each time. By construction, any occurrence of the fault is guaranteed to be detected since

it renders (P_{MI}) infeasible and a false alarm never occurs. In case, T -detectability cannot be verified for a finite T , one can still use (P_{MI}) in a receding horizon manner but the occurrence of the fault can be missed in this case unless one uses increasing amounts of data at each step. It is also worth noting that if an unmodeled fault occurs, it can still be detected in this manner; i.e., any detection corresponds to an anomaly in the behavior of the system.

5 Generalizations

Thanks to the generality of the modeling framework, the fault detection scheme presented in this paper can be applied to several other problems without any modifications. We briefly mention some of these problems in this section.

5.1 Attack Detection

Recently, there has been a considerable interest in attack detection [48] and attack resilient estimation [49, 50] for cyber-physical systems. One can argue that fault detection and attack detection are different problems since, in general, an attacker, in addition to harming the system, has the intention of not being noticed whereas the latter is not a concern for faults. However, since the proposed framework takes a worst-case approach, it is also well suited for attack detection. The only slight difference is that attack models are typically subject to more uncertainty than fault models.

Consider the following linear state-space model subject to sensor attack taken from [49]:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{a}_k + \boldsymbol{\eta}_k \end{aligned} \quad (33)$$

where $\mathbf{a}_k \in \mathbb{R}^{n_y}$ is the attack vector at time instance k .

The system model corresponds to the case where $\mathbf{a}_k = 0$ in (33). Take as the fault (i.e., attack) model as a switched linear system with $\bar{s} = \sum_{i \in \mathbb{Z}_a^+} \binom{n_y}{i}$ modes, where each mode represents one possible attack scenario where at most a sensors are attacked. If $\mathcal{A} \subseteq \mathbb{Z}_{n_y}^+$, $|\mathcal{A}| \leq a$, is the set of sensors attacked in mode j , then mode j has the form (33), with uncertain parameters $\mathbf{a}_k^i \neq 0$ (or, $|\mathbf{a}_k^i| \geq \epsilon$ if a minimum attack strength ϵ is known), $i \in \mathcal{A}$. Conditions like “the same set of sensors are attacked persistently” can be included using indicators that restrict the switching between attack modes. If this attack model can be shown to be T -detectable, running model invalidation

problem in a receding horizon manner detects the attacks. Moreover, if additional observability conditions hold such as s -sparse observability in [49] with $s = 2a$, model invalidation problem can be used for resilient state estimation with an additional constraint that limits the number of attacked sensors. One can also precisely estimate the trade-off between the noise level and attack magnitude by solving T -detectability problems with appropriately defined objectives. Note that our framework trivially extends to the case where the system itself is switched or there are other types of attacks if the goal is detecting the attacks. Obtaining conditions for state estimation guarantees in this general setting is left for future work.

5.2 Cascaded Faults

By using indicators and the fact that the proposed framework handles switched systems, it is possible to relax Assumption 1 on the persistency of the faults. In general, a system is subject to multiple faults and these faults can occur in a cascaded fashion. Assume each such fault is represented by an SWA model. It is possible to define a new SWA model whose number of modes is the sum of the number of modes of the system model and each fault model. Then, the cascading faults can be represented by restricting the mode sequences of this new SWA model with indicators that capture potential a priori knowledge on the order of failures. Therefore, one can apply the proposed framework to analyze T -detectability of multiple cascading faults and to detect non-permanent faults.

6 Illustrative Examples

In this section, we consider one set of numerical examples and an example motivated by building radiant systems to illustrate the efficacy of the methods proposed in this paper. All the examples are implemented on a 3.5 GHz machine with 32 GB of memory running Ubuntu. For the implementation of model invalidation approach and finding T for T -detectability, we utilized Yalmip [51] and CPLEX [9]. All the approaches and examples are implemented in Matlab, and are available with MI4Hybrid² toolbox.

6.1 Numerical Results

In this section, we illustrate the efficiency of the model invalidation approach proposed in this paper in terms of execution-time for various time-horizons

²<https://github.com/data-dynamics/MI4Hybrid>

and number of modes of the system. Consider a hidden-mode switched affine model, \mathcal{G}_6 , with admissible sets $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\| \leq 11\}$, $\mathcal{U} = \{\mathbf{u} \mid \|\mathbf{u}\| \leq 1000\}$ and $\mathcal{E} = \{\boldsymbol{\eta} \mid \|\boldsymbol{\eta}\| \leq 0.1\}$. We assume there is no process noise and no uncertainty in the parameters of the system. \mathcal{G}_6 has three states and six modes. We assume a fixed $B = [1 \ 0 \ 1]^T$ and $C = [1 \ 1 \ 1]$ for all modes. The system matrices of the modes are:

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} 0.5 & 0.5 & 0.5 \\ 0.1 & -0.2 & 0.5 \\ -0.4 & 0.6 & 0.2 \end{pmatrix}, \mathbf{f}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{A}_2 = \begin{pmatrix} 0.5 & 0.5 & 0.5 \\ -0.3 & -0.2 & 0.3 \\ 0.1 & -0.3 & -0.5 \end{pmatrix}, \mathbf{f}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ \mathbf{A}_3 &= \begin{pmatrix} 0.5 & 0.2 & 0.6 \\ 0.2 & -0.2 & 0.2 \\ -0.9 & 0.7 & 0.1 \end{pmatrix}, \mathbf{f}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \mathbf{A}_4 = \begin{pmatrix} -0.5 & 0.5 & 0.8 \\ 0.1 & -0.2 & -0.6 \\ 0.2 & -0.6 & 0.3 \end{pmatrix}, \mathbf{f}_4 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \\ \mathbf{A}_5 &= \begin{pmatrix} 0.8 & 0.5 & 0.2 \\ -0.1 & 0.2 & -0.3 \\ 0.5 & 0.4 & -0.1 \end{pmatrix}, \mathbf{f}_5 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \mathbf{A}_6 = \begin{pmatrix} -0.3 & 0.8 & -0.1 \\ 0.4 & -0.1 & 0.3 \\ 0.9 & -0.2 & 0.6 \end{pmatrix}, \mathbf{f}_6 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Let us define, $\mathcal{G}_t, t = 1, \dots, 6$, to be a system with the first t modes of \mathcal{G}_6 . In addition, suppose the fault, \mathcal{G}^f , is an affine model with the following system matrices:

$$\mathbf{A}^f = \begin{pmatrix} 0.8 & 0.7 & 0.6 \\ 0.1 & -0.2 & 0.3 \\ -0.4 & 0.3 & -0.2 \end{pmatrix}, \mathbf{B}^f = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{f}^f = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \quad (34)$$

We first generate input-output trajectories from \mathcal{G}^f of various lengths. We then fix the length at 100 samples and investigate the effect of increasing the number of modes of the system. Both examples are repeated for 20 times for randomly generated input and noise sequences. The results for applying model invalidation in this setting are illustrated in Fig. 2. Although the execution-time scales relatively reasonably with increasing horizon and number of modes, depending on the time-scale of the system dynamics, solving only short horizon problems might be feasible for real-time implementations, motivating T -detectability.

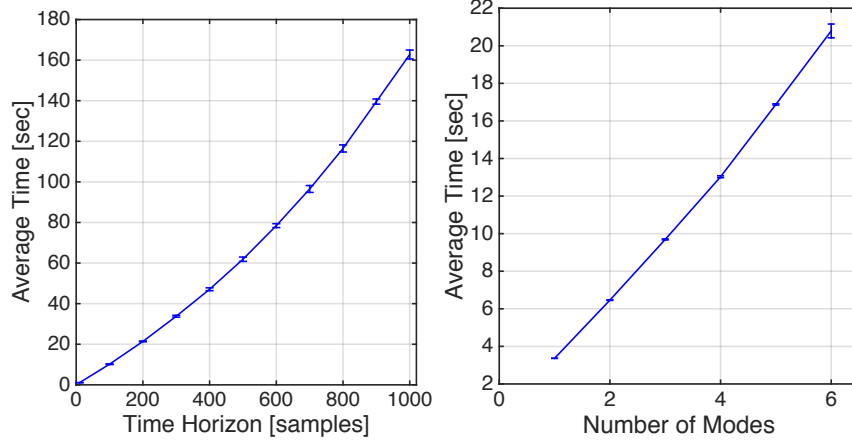


Figure 2: The average execution time for invalidation of data generated by \mathcal{G}^f for \mathcal{G}_3 on various time horizons (left), and for \mathcal{G}_t on a fixed time horizon of length 100 for different values of t (right).

6.2 Building Radiant Systems

6.2.1 System Model

We consider a building with four rooms, which has a radiant system with two pumps, adapted from [52] and illustrated in Fig. 3. We assume that the two pumps can either be on with known constant flow or off. Each pump is connected to a valve, which adjusts the constant flow of the pump. The overall system can be described by a SWA model with six states associated with the temperature of the water in the two cores and the temperature of the four rooms. We assume all the states are measured with some measurement noise. The system has four modes: mode 1 corresponds to the case where both pumps are off, mode 2 is active when pump 1 is on and pump 2 is off, mode 3 represents the case when pump 1 is off and pump 2 is on, and mode 4 is when both pumps are on. When both pumps are on the core water temperatures, $T_{c,1}, T_{c,2}$ evolve as a function of the parameters of the system and the temperatures of the zones, $T_i, i \in \mathbb{Z}_4^+$, according to:

$$C_{r,1}\dot{T}_{c,1}(t) = K_{r,1}(T_1 - T_{c,1}) + K_{r,3}(T_3 - T_{c,1}) + K_{w,1}(T_{w,1} - T_{c,1}) \quad (35)$$

$$C_{r,2}\dot{T}_{c,2}(t) = K_{r,2}(T_2 - T_{c,2}) + K_{r,4}(T_4 - T_{c,2}) + K_{w,2}(T_{w,2} - T_{c,2}). \quad (36)$$

If the first (second) pump is off, the last term in (35) (in (36)) becomes zero. The rest of state equations do not change with the status of the two pumps

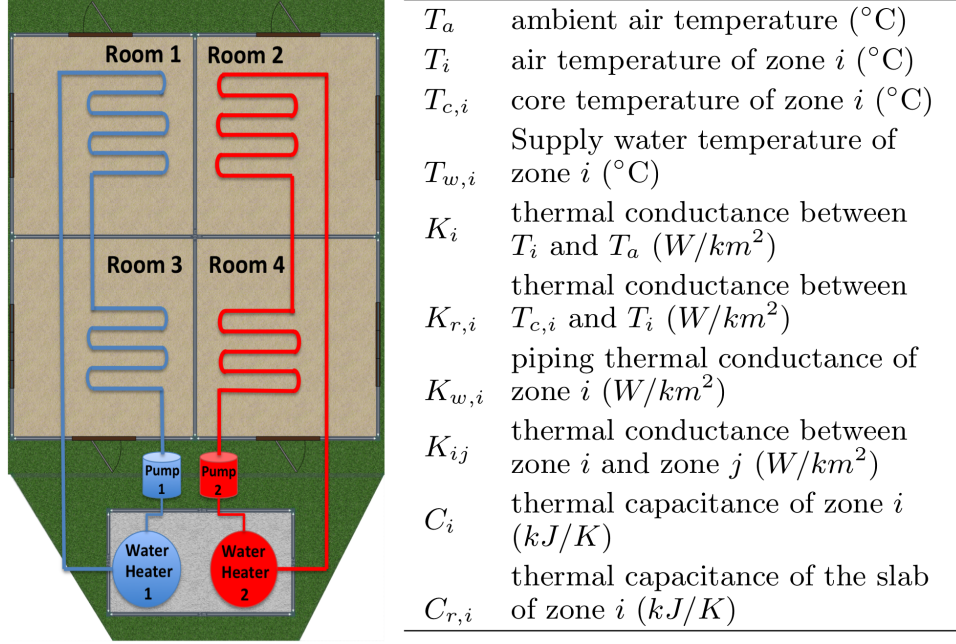


Figure 3: Left: four zone radiant system with two pumps. Right: parameter definitions and units for the radiant system.

and are as follows:

$$\begin{aligned}
 C_1 \dot{T}_1(t) &= K_{r,1}(T_{c,1} - T_1) + K_1(T_a - T_1) + K_{12}(T_2 - T_1) + K_{13}(T_3 - T_1) \\
 C_2 \dot{T}_2(t) &= K_{r,2}(T_{c,2} - T_2) + K_2(T_a - T_2) + K_{12}(T_1 - T_2) + K_{24}(T_2 - T_4) \\
 C_3 \dot{T}_3(t) &= K_{r,1}(T_{c,1} - T_3) + K_3(T_a - T_3) + K_{13}(T_1 - T_3) + K_{34}(T_4 - T_3) \\
 C_4 \dot{T}_4(t) &= K_{r,2}(T_{c,2} - T_4) + K_4(T_a - T_4) + K_{24}(T_2 - T_4) + K_{34}(T_3 - T_4).
 \end{aligned}$$

The list of parameters are given in Fig. 3 and their values³ are chosen according to the ranges provided in [52].

The discrete-time switched affine model $\mathcal{G}_R = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{\mathcal{G}_i^\Delta\}_{i=1}^4)$ for the system is obtained with sampling time of 5 minutes, where $\mathcal{X} = \{x \mid 15 \leq x_i \leq 19\}$, $\mathcal{U} = \emptyset$, $\mathcal{E} = \{\boldsymbol{\eta} \mid \|\boldsymbol{\eta}\| \leq 0.05\}$, and G_i is the discrete uncertain

³ $T_a = 10$, $K_1 = \frac{1}{2.1}$, $K_2 = \frac{1}{2.1}$, $K_3 = \frac{1}{2.2}$, $K_4 = \frac{1}{2.2}$, $K_{r,1} = \frac{1}{0.125}$, $K_{r,2} = \frac{1}{0.130}$, $K_{r,3} = \frac{1}{0.125}$, $K_{r,4} = \frac{1}{0.130}$, $K_{12} = \frac{1}{0.16}$, $K_{13} = \frac{1}{0.16}$, $K_{24} = \frac{1}{0.16}$, $K_{34} = \frac{1}{0.16}$, $K_{w,1} = \frac{1}{0.07}$, $K_{w,2} = \frac{1}{0.05}$, $C_1 = 1900$, $C_2 = 2100$, $C_3 = 2000$, $C_4 = 1800$, $C_{r,1} = 3000$, $C_{r,2} = 4000$, $T_{w,1} = 18$, $T_{w,2} = 18$.

affine model of the i th mode. We assume the uncertainty in the system is due to small changes in ambient temperature ($T_a = 10 + \delta$ with $\|\delta\| \leq 0.5$). The effect of the uncertainty appears in the \mathbf{f}_i vector for each mode i .

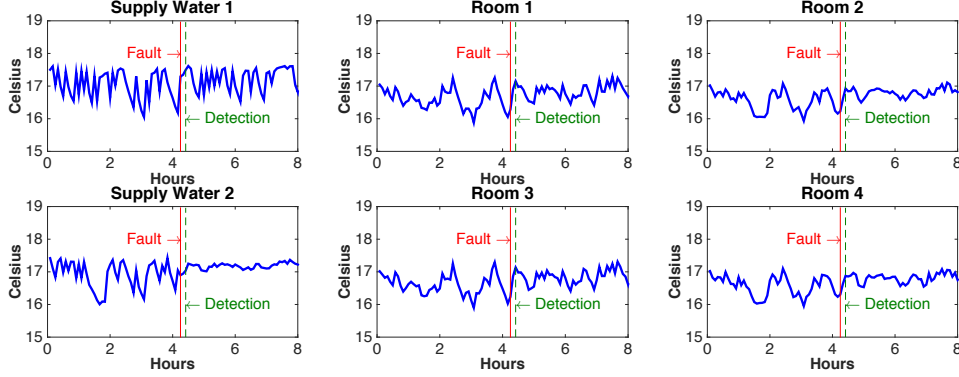


Figure 4: Fault detection results on the data, which consists of outputs of the system until time sample 50 and the outputs of the faulty system afterwards.

6.2.2 Fault model

We assume that the valve of the second pump stuck in middle and do not respond to on or off commands. In faulty mode the system has only two modes, which correspond to the “on” or “off” states of the first pump. This fault is modeled with a change in the second heat conductance parameter, which is assumed to be $K_{w,2} = \frac{0.5}{0.05}$ when the fault occurs. This indicates the fact that the heat transfer is cut in half, because of the slower water flow. We assume the faulty model is represented by $\mathcal{G}_R^f = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{\mathcal{G}_i^{\Delta, f}\}_{i=1}^2)$. This is an incipient fault and hard to detect because the outputs can remain within the reasonable range, and do not change dramatically.

6.2.3 Results

T -detectability: First, we analyze the T -detectability of the pair $(\mathcal{G}_R, \mathcal{G}_R^f)$. In particular, our approach finds that for $T = 8$, this fault is detectable for the radiant system. The iterations from $T = 1$ until detectability is verified at $T = 8$ took 36.57 seconds with the proposed approach. As a comparison, satisfiability modulo theory (SMT) based approach proposed in [10] took 10008.19 seconds. Although finding a T such that a particular fault is T -detectable for a specific system is an offline step, hence, one can tolerate a

slightly slower algorithm, the MILP based approach proposed in this paper shows significant improvement in execution time.

Fault detection: Next, we demonstrate the proposed fault detection scheme on this system. In this simulation, we generate data from the radiant system model \mathcal{G}_R for 50 samples (four hours and ten minutes), and the persistent fault \mathcal{G}_R^f becomes active at time sample 51. The model invalidation problem is solved in receding horizon manner with horizon size 8. The fault is detected at time sample 53. Simulation traces are shown in Fig. 4. It is worth mentioning that T is calculated based on the worst case scenario to provide guarantees for detection, but for a particular realization of input-output trajectory usually it is possible to detect the fault earlier as is the case in this example.

Redundant sensors: In this subsection, we show how T -detectability analysis can be used for selecting “optimal” sensors for fault detection. We consider five different scenarios, each corresponding to different set of states being measured as shown in Tab. 3. For each of the scenarios we consider the case with no uncertainty and noise in the system and fault models, e.g., $\mathcal{E} = \{\mathbf{0}\}$ and $\Omega = \{\mathbf{0}\}$, where $\mathbf{0}$ indicates a vector of zeros of appropriate dimension; and also the noise and uncertainty models used in this section earlier. We compute the minimum T -values for each scenario to achieve T -detectability as listed in Tab. 3. From these results we see that measuring the core temperatures seems crucial, especially in case of uncertainty and noise, for detecting this fault.

Table 3: Value of T for the five scenarios, under two model assumptions of with and without uncertainty and noise.

Measured States	T (no uncertainty)	T (with uncertainty)
$T_{c,1}, T_{c,2}, T_1, T_2, T_3, T_4$	1	6
$T_{c,2}, T_1, T_2, T_3, T_4$	1	6
T_1, T_2, T_3, T_4	2	> 100
$T_{c,1}, T_{c,2}, T_1, T_2$	2	6
$T_{c,2}, T_1, T_3$	2	7

Weak-detectability: Consider a variant of the fault model described in Subsection 6.2.2 where the second valve can be closed, but when the pump is on it can only open up to the half of its capacity. Such a fault is not T -detectable for any finite T , because it shares two modes with the a priori system model. Problem (P_T) is always feasible, because there always exists a switching sequence that matches the fault and system model, which is the one that keeps the second pump always off. In order to detect such a fault, activation of the “on” mode for second pump is necessary, which corresponds to the third and fourth modes of the fault model. We capture this with an indicator of the form $\mathcal{I}_t = (\{3, 4\}, 1, 1, =)$. This indicator is incorporated in to problem (P_T) via the constraint: $\sum_{i \in \mathbb{Z}_4^+} \sum_{j=3}^4 d_{i,j,t-T+1} = 1$. Solving problem (P_T) with this extra constraint renders $T = 8$. In order to illustrate the use of weak-detectability in the fault detection scheme, we generate an output sequence from the faulty model as follows: (i) samples 1-30 are generated by the first two modes of faulty system; (ii) sample 31 is generated by one of the last two modes of the faulty system, and is followed by 9 samples generated by the first two modes; and (iii) samples 41-75 are generated by allowing any mode of the faulty system to be active. As illustrated in Fig. 5, the first detection occurs immediately after the first switch to the modes where second pump is “on”. The fault is also detected for all samples after 43, where there is switching to the last two modes. Recall that, the mode signal is not measured by our fault detection scheme. As one can see the fault is not detected in the first 30 samples, because the second pump is always “off” during that period. It is worth mentioning that the fault does not affect the behavior of the system in the first two hours and forty minutes and detecting it is impossible.

7 Conclusions and Discussion

In this paper, we present a fault detection scheme that guarantees the detection of particular faults and can be implemented in real-time for many applications. The proposed scheme is applicable to an expressive class of system and fault models, namely hidden-mode switched affine models with parametric uncertainty. The modeling framework is further extended to include language constraints on the modes and necessary and sufficient conditions for detectability using a receding horizon scheme are provided based on MILP. The key step in efficiency of implementation is the ability to preserve detection guarantees while using the receding horizon scheme as a result of T -detectability property. T -detectability essentially tells us how much data is enough for detecting a fault by utilizing the knowledge of a

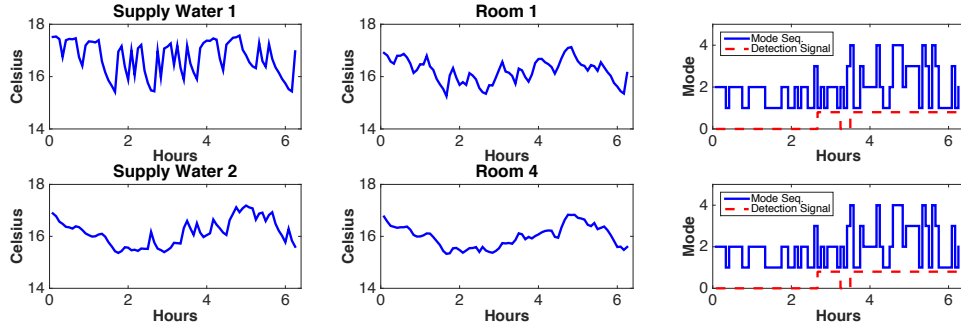


Figure 5: Weak-detectability results, where one of the last two modes of faulty model is activated at sample 31 for the first time and after sample 41. The red solid line in the bottom figures is 1 when a fault is detected.

(possibly uncertain) fault model and the structure in the dynamics. With the advances in MILP-solvers, there has been a resurgence of interest in using it in control design and real-time implementations [53]. Here we leverage such advances in the context of fault detection. In particular, since we only solve MILP feasibility problems (as opposed to optimization), this leads to a fairly efficient scheme as demonstrated via examples.

In the future, we are interested in applying similar ideas to fault isolation and active fault detection. We have also some ongoing work on fault detection for nonlinear hybrid systems.

References

- [1] V. Rajah. Taming the data deluge, 2014.
- [2] M. Sznaier, O. Camps, N. Ozay, and C. Lagoa. Surviving the upcoming data deluge: A systems and control perspective. In *IEEE CDC*, Dec 2014.
- [3] J. Weimer, J. Araujo, M. Amoozadeh, S. Ahmadi, H. Sandberg, and K. Johansson. Parameter-invariant actuator fault diagnostics in cyber-physical systems with application to building automation. In *Control of CPS*, pages 179–196. Springer International Publishing, 2013.
- [4] R. Burkart, K. Margellos, and J. Lygeros. Nonlinear control of wind turbines: An approach based on switched linear systems and feedback linearization. In *IEEE CDC-ECC*, pages 5485–5490, 2011.

- [5] Z. Sun. *Switched linear systems: control and design*. Springer Sc. & Bus. Med., 2006.
- [6] P. Rosa, C. Silvestre, J. Shamma, and M. Athans. Fault detection and isolation of LTV systems using set-valued observers. In *IEEE CDC*, pages 768–773, 2010.
- [7] N. Ozay, M. Sznaier, and C. Lagoa. Model (in)validation of switched ARX systems with unknown switches and its application to activity monitoring. In *IEEE CDC*, pages 7624–7630, 2010.
- [8] N. Ozay, M. Sznaier, and C. Lagoa. Convex certificates for model (in)validation of switched affine systems with unknown switches. *IEEE Trans. Autom. Control*, 59(11):2921–2932, Nov 2014.
- [9] IBM ILOG CPLEX. User’s manual for CPLEX. *Int. Bus. Mach. Corp.*, 46(53):157, 2009.
- [10] F. Harirchi and N. Ozay. Model invalidation for switched affine systems with applications to fault and anomaly detection. *IFAC ADHS*, 48(27):260–266, 2015.
- [11] R. Beard. *Failure accommodation in linear systems through self-reorganization*. PhD thesis, MIT, 1971.
- [12] H. Jones. *Failure detection in linear systems*. PhD thesis, MIT, 1973.
- [13] S. Simani, C. Fantuzzi, and R. Patton. *Model-based fault diagnosis in dynamic systems using identification techniques*. Springer Sc. & Bus. Med., 2003.
- [14] R. Isermann. *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Sc. & Bus. Med., 2006.
- [15] S. Ding. *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Sc. & Bus. Med., 2008.
- [16] R. Patton, P. Frank, and R. Clark. *Issues of fault diagnosis for dynamic systems*. Springer Sc. & Bus. Med., 2013.
- [17] R. Isermann. Fault diagnosis of machines via parameter estimation and knowledge processing—tutorial paper. *Automatica*, 29(4):815–835, 1993.
- [18] P. Frank. Advances in observer-based fault diagnosis. In *Int. Conf. on Fault Diag.: TOOLDIAG*, 1993.

- [19] R. Patton and J. Chen. Observer-based fault detection and isolation: robustness and applications. *Cont. Eng. Prac.*, 5(5):671–682, 1997.
- [20] I. Shames, A. Teixeira, H. Sandberg, and K. Johansson. Distributed fault detection for interconnected second-order systems. *Automatica*, 47(12):2757–2764, 2011.
- [21] J. Gertler. Fault detection and isolation using parity relations. *Cont. Eng. Prac.*, 5(5):653–661, 1997.
- [22] P. Rosa and C. Silvestre. Fault detection and isolation of LPV systems using set-valued observers: An application to a fixed-wing aircraft. *Cont. Eng. Prac.*, 21(3):242–252, 2013.
- [23] R. Nikoukhah. Guaranteed active failure detection and isolation for linear dynamical systems. *Automatica*, 34(11):1345–1358, 1998.
- [24] R. Nikoukhah and S. Campbell. Auxiliary signal design for active failure detection in uncertain linear systems with a priori information. *Automatica*, 42(2):219–228, 2006.
- [25] J. K. Scott, R. Findeisen, R. D Braatz, and D. M. Raimondo. Input design for guaranteed fault diagnosis using zonotopes. *Automatica*, 50(6):1580–1589, 2014.
- [26] E. Garcia and P. Frank. Deterministic nonlinear observer-based approaches to fault diagnosis: a survey. *Cont. Eng. Prac.*, 5(5):663–670, 1997.
- [27] H. Hammouri, M. Kinnaert, and E. El Yaagoubi. Observer-based approach to fault detection and isolation for nonlinear systems. *IEEE Trans. Autom. Control*, 44(10):1879–1884, 1999.
- [28] W. Pan, Y. Yuan, H. Sandberg, J. Gonçalves, and G. Stan. Online fault diagnosis for nonlinear power systems. *Automatica*, 55:27–36, 2015.
- [29] C. De Persis and A. Isidori. A geometric approach to nonlinear fault detection and isolation. *IEEE Trans. Autom. Control*, 46(6):853–865, 2001.
- [30] S. McIlraith, G. Biswas, D. Clancy, and V. Gupta. Hybrid systems diagnosis. In *Hybrid Systems: Computation and Control*, pages 282–295. Springer, 2000.

- [31] S. Narasimhan and G. Biswas. Model-based diagnosis of hybrid systems. *IEEE Trans. Syst., Man, Cybern. A*, 37(3):348–361, 2007.
- [32] Y. Deng, A. D’Innocenzo, M. Di Benedetto, S. Di Gennaro, and A. Julius. Verification of hybrid automata diagnosability with measurement uncertainty. *IEEE Trans. Autom. Control*, 61(4):982–993, 2016.
- [33] M. Grewal and K. Glover. Identifiability of linear and nonlinear dynamical systems. *IEEE Trans. Autom. Control*, 21(6):833–837, 1976.
- [34] V. René, A. Chiuso, and S. Soatto. Observability and identifiability of jump linear systems. In *IEEE CDC*, volume 4, pages 3614–3619, 2002.
- [35] M. Babaali and M. Egerstedt. Observability of switched linear systems. In *Int. Workshop on Hybrid Systems: Computation and Control*, pages 48–63. Springer, 2004.
- [36] H. Lou and P. Si. The distinguishability of linear control systems. *Nonlinear Analysis: Hybrid Systems*, 3(1):21–38, 2009.
- [37] P. Rosa and C. Silvestre. On the distinguishability of discrete linear time-invariant dynamic systems. In *IEEE CDC-ECC*, pages 3356–3361, 2011.
- [38] N. Adnan, I. Izadi, and T. Chen. On expected detection delays for alarm systems with deadbands and delay-timers. *Journal of Process Control*, 21(9):1318–1331, 2011.
- [39] M. Mariton. Detection delays, false alarm rates and the reconfiguration of control systems. *International Journal of Control*, 49(3):981–992, 1989.
- [40] A. Stoorvogel, H. Niemann, and A. Saberi. Delays in fault detection and isolation. In *ACC*, volume 1, pages 459–463, 2001.
- [41] R. Smith and J. Doyle. Model validation: A connection between robust control and identification. *IEEE Trans. Autom. Control*, 37(7):942–952, 1992.
- [42] Y. Cheng, Y. Wang, M. Sznaier, N. Ozay, and C. Lagoa. A convex optimization approach to model (in)validation of switched arx systems with unknown switches. In *IEEE CDC*, pages 6284–6290, Dec 2012.

- [43] F. Harirchi, Luo Z., and N. Ozay. Model (in)validation and fault detection for systems with polynomial state-space models. In *ACC*, pages 1017–1023, 2016.
- [44] R. Raman and I. Grossmann. Modelling and computational techniques for logic based integer programming. *Comput. & Chem. Eng.*, 18(7):563–578, 1994.
- [45] D. Bertsimas and M. Sim. Tractable approximations to robust conic optimization problems. *Math. Prog.*, 107(1-2):5–36, 2006.
- [46] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, and D. Teneketzis. Diagnosability of discrete-event systems. *IEEE Trans. Autom. Control*, 40(9):1555–1575, 1995.
- [47] C. Baier and J. Katoen. *Principles of model checking*, volume 26202649. MIT press Cambridge, 2008.
- [48] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Trans. Autom. Control*, 58(11):2715–2729, 2013.
- [49] Y. Shoukry, P. Nuzzo, A. Puggelli, A. Sangiovanni-Vincentelli, S. Seshia, and P. Tabuada. Secure state estimation for cyber physical systems under sensor attacks: a satisfiability modulo theory approach. *arXiv preprint arXiv:1412.4324*, 2014.
- [50] M. Chong, M. Wakaiki, and J. Hespanha. Observability of linear systems under adversarial attacks. In *ACC*, pages 2439–2444, 2015.
- [51] J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *CACSD Conference*, Taipei, Taiwan, 2004.
- [52] T. Nghiem, G. Pappas, and R. Mangharam. Event-based green scheduling of radiant systems in buildings. In *ACC*, pages 455–460, 2013.
- [53] V. Raman, A. Donzé, D. Sadigh, R. M. Murray, and S. A. Seshia. Reactive synthesis from signal temporal logic specifications. In *Hybrid Systems: Computation and Control*, pages 239–248. ACM, 2015.

A Appendix - Converting absolute value constraints to mixed integer linear constraints:

Consider the first constraint in (17c):

$$|\delta_{i,k}^{A^{m_l}}| \leq |[\mathbf{A}_i^N]^{m,l}| |\mathbf{z}_{i,k}^l|$$

If $|[\mathbf{A}_i^N]^{m,l}|$ is zero, then $\delta_{i,k}^{A^{m_l}} = 0$ and both the new variable and constraint can be removed. Otherwise, it is a positive constant, and we can divide both sides of inequality by that and define new variables as follows:

$$|\sigma_{i,k}^{A^{m_l}}| \leq |\mathbf{z}_{i,k}^l|$$

There are four different cases that are possible, and if any of them occurs, then the constraint above is satisfied. In fact only one of the following four cases can occur at each time, so the constraint above is equivalent to the logical OR of the following four cases:

- $\sigma_{i,k}^{A^{m_l}} \geq 0, \mathbf{z}_{i,k}^l \geq 0, \sigma_{i,k}^{A^{m_l}} \leq \mathbf{z}_{i,k}^l$
- $\sigma_{i,k}^{A^{m_l}} \leq 0, \mathbf{z}_{i,k}^l \leq 0, \sigma_{i,k}^{A^{m_l}} \geq \mathbf{z}_{i,k}^l$
- $\sigma_{i,k}^{A^{m_l}} \geq 0, \mathbf{z}_{i,k}^l \leq 0, \sigma_{i,k}^{A^{m_l}} \leq -\mathbf{z}_{i,k}^l$
- $\sigma_{i,k}^{A^{m_l}} \leq 0, \mathbf{z}_{i,k}^l \geq 0, \sigma_{i,k}^{A^{m_l}} \geq -\mathbf{z}_{i,k}^l$

In order to implement the logical OR between the four cases, we need to introduce four binary variables: b_1, b_2, b_3, b_4 . The equivalent MIP constraints are as follows:

$$b_1 \sigma_{i,k}^{A^{m_l}} \geq 0, b_1 \mathbf{z}_{i,k}^l \geq 0, b_1 (\sigma_{i,k}^{A^{m_l}} - \mathbf{z}_{i,k}^l) \leq 0 \quad (37)$$

$$b_2 \sigma_{i,k}^{A^{m_l}} \leq 0, b_2 \mathbf{z}_{i,k}^l \leq 0, b_2 (\sigma_{i,k}^{A^{m_l}} - \mathbf{z}_{i,k}^l) \geq 0 \quad (38)$$

$$b_3 \sigma_{i,k}^{A^{m_l}} \geq 0, b_3 \mathbf{z}_{i,k}^l \leq 0, b_3 (\sigma_{i,k}^{A^{m_l}} + \mathbf{z}_{i,k}^l) \leq 0 \quad (39)$$

$$b_4 \sigma_{i,k}^{A^{m_l}} \leq 0, b_4 \mathbf{z}_{i,k}^l \geq 0, b_4 (\sigma_{i,k}^{A^{m_l}} + \mathbf{z}_{i,k}^l) \geq 0 \quad (40)$$

with $\sum_{i \in \mathbb{Z}_4^+} b_i = 1$. Finally, in order to transform these constraints to MILP, we define $\tilde{\mathbf{z}}_{i,q,k}^l = b_q \mathbf{z}_{i,k}^l$, $\sigma_{i,q,k}^{A^{m_l}} = b_q \sigma_{i,k}^{A^{m_l}}$ and $\tilde{a}_{i,q,k} = a_{i,k} b_q$. Therefore, $\sum_{q \in \mathbb{Z}_4^+} \tilde{\mathbf{z}}_{i,q,k}^l = \mathbf{z}_{i,k}^l$ and $\sum_{q \in \mathbb{Z}_4^+} \tilde{a}_{i,q,k} = a_{i,k}$. The following constraints ensure that $\sigma_{i,q,k}^{A^{m_l}}, \tilde{\mathbf{z}}_{i,q,k}^l$ are only nonzero for one of the four cases, without adding any extra constraints.

$$\tilde{a}_{i,q,k} X_l \leq \sigma_{i,q,k}^{A^{m_l}} \leq \tilde{a}_{i,q,k} X_u$$

$$\tilde{a}_{i,q,k} X_l \leq \tilde{\mathbf{z}}_{i,q,k}^l \leq \tilde{a}_{i,q,k} X_u.$$